Charotar University of Science and Technology
Chandubhai s. patel institute of Technology
B. Tech CE, CSE, IT & AIML
Subject: OCAIML4001: Responsible & Safe AI Systems
Practice Set

Sr no.	Questions	Marks	BL	Cos
1.	A global healthcare platform trains a large-scale AI model on patient records from diverse populations. Recently, the model was found to underperform for certain minority groups. Identify three imminent risks associated with such large-scale AI deployments, and justify how each risk could negatively affect patient outcomes.	3	AN	1,2
2.	Consider a self-driving delivery robot assigned to "deliver as fast as possible." Without adjusting its directive, it begins violating traffic rules. Analyze how such a scenario reflects goal misspecification, and propose a strategy to align AI Behavior with human values.	2	AP	1,2
3.	A leading AI research firm publishes a breakthrough in general-purpose agents. However, concerns emerge around malicious actors using it to generate fake news, and the agent behaving unpredictably in unsupervised environments. Assess this situation by classifying it under the long-term risks of Artificial Intelligence—misuse**,** misgeneralization**,** and rogue Artificial General Intelligence (AGI)—with appropriate examples.	4	E	1,2
4.	As a policymaker designing national AI guidelines, you are asked to embed Responsible AI principles to guide public sector deployments. Give any four key principles and suggest actionable measures to effectively implement each in practice.	4	AN	1,2
5.	A computer vision system deployed in a retail store is tricked by adversarial stickers placed on items, misidentifying products at checkout. Contrast how such adversarial attacks manifest differently in vision-based models compared to natural language processing (NLP) systems, citing real-world or research-based examples.	4	AN	1,2
6.	A tech company mandates that all deployed AI systems must be interpretable to end-users. As a lead engineer, how would you integrate interpretability features into a vision or language model, and which two recent advancements in either domain would you draw upon to improve responsible deployment?	3	AP	1,3

7	A financial fraud detection system is compromised through a poisoning attack. In what way does a Trojan attack alter the functioning of the artificial intelligence system? Provide a relevant real-world scenario.	4	AP	3,4
8	A healthcare organization uses a machine learning system to detect chronic diseases from patient records. Discuss the role of mechanistic interpretability and representation engineering in improving transparency in this case, including their potential benefits and limitations for artificial intelligence safety. 1. In what ways can mechanistic interpretability techniques help reveal whether the model is relying on medically valid features rather than spurious correlations? Illustrate with an example (4 Marks). 2. Evaluate representation engineering for reducing demographic bias while balancing accuracy and fairness in healthcare AI systems (3 Marks).	7	AN	3,4
9	An educational technology company uses a recommendation system for students. Parents and teachers raise concerns about fairness in recommendations and protection of personal data. Examine the role of privacy-preserving techniques and fairness metrics in addressing these concerns.	5	E	5,6
10	An autonomous driving company aims to improve trust in its system for tasks like emergency braking and pedestrian recognition. Propose a framework to enhance transparency and identify the key components of this framework.	4	C	3,4
11	An AI model used for credit scoring shows excellent accuracy but offers no clear reasoning for its decisions. Regulators demand interpretability reports. Discuss how mechanistic interpretability and representation engineering can help reveal inner workings of the model, and why this is crucial for AI safety.			4
12	A global social media company introduces an AI moderation system to flag toxic or misleading content. After deployment, users report censorship of harmless posts and bias toward specific languages. Analyze the ethical and fairness implications of such AI moderation systems and propose two strategies to improve inclusiveness. Analyze (AN)	5	AN	1,2
13	A bank deploys a credit approval model that unintentionally denies loans to applicants from low-income neighborhoods. Identify two fairness metrics that could be used to detect this bias, and explain how each helps ensure fair decision-making.	5	AP	2,5
14	An AI-based hiring platform uses candidate résumés for automated screening. Later, it's discovered that the system favors male candidates for technical roles. Evaluate the role of bias audits and fairness-enhancing interventions to correct such discriminatory behavior.	5	E	2,5
15	A hospital uses federated learning to train AI models on patient data from multiple clinics without sharing raw records. Explain how privacy-preserving techniques like differential privacy or secure aggregation contribute to ethical AI in healthcare.	5	AP	1,5

16	As a policymaker designing national AI guidelines, you are asked to embed Responsible AI principles to guide public sector deployments. Give any four key principles and suggest actionable measures to effectively implement each in practice.	5	AN	1,2

S. No.	Question	Marks	BL	Cos
17	A global social media company introduces an AI moderation system to flag toxic or misleading content. After deployment, users report censorship of harmless posts and bias toward specific languages. Analyze the ethical and fairness implications of such AI moderation systems and propose two strategies to improve inclusiveness.	5	AN	1,2
18	A bank deploys a credit approval model that unintentionally denies loans to applicants from low-income neighborhoods. Identify two fairness metrics that could be used to detect this bias, and explain how each helps ensure fair decision-making.	5	AP	2, 5
19	An AI-based hiring platform uses candidate résumés for automated screening. Later, it's discovered that the system favors male candidates for technical roles. Evaluate the role of bias audits and fairness-enhancing interventions to correct such discriminatory behavior.	5	E	2, 5
20	A hospital uses federated learning to train AI models on patient data from multiple clinics without sharing raw records. Explain how privacy-preserving techniques like differential privacy or secure aggregation contribute to ethical AI in healthcare.	5	AP	1, 5
21	A government agency plans to adopt AI-based surveillance for urban safety. Critically discuss how privacy, accountability, and proportionality should be embedded in such systems to comply with Responsible AI principles.	5	E	1, 6
22	During an audit, an AI insurance model is found to have strong predictive performance but lacks transparency in decision criteria. Propose a set of explainability tools (such as LIME, SHAP, or GradCAM) and describe how each could enhance model accountability.	5	C	3, 4
23	An e-commerce recommendation system starts suggesting unsafe or age-inappropriate products to minors. Analyze how adversarial testing and explainability mechanisms could be used to detect and prevent such failures.	5	AN	3, 4
24	India's DPDP Act (2023) and the EU's GDPR both regulate data processing by AI systems**.** Compare their approaches to user consent, data storage, and accountability, and explain how these frameworks promote AI safety.	5	AN	6
25	The European Union's AI Act categorizes systems as "unacceptable risk," "high risk," or "limited risk." Explain the rationale behind this classification and evaluate how it could guide ethical AI deployment in India.	5	E	6
26	A university introduces an AI grading tool for student essays. Soon, concerns arise about bias against non-native English speakers. Discuss how explainability reports and participatory design can be used to rebuild stakeholder trust.	5	AP	5, 7
27	A law enforcement agency adopts a predictive policing model to forecast crime-prone areas. Analyze the potential ethical, legal, and social risks, and suggest guidelines for responsible use of such models.	5	AN	1, 6
28	A health-tech startup develops an AI model that suggests personalized treatments based on genetic data. Evaluate the implications for privacy, informed consent, and fairness when using sensitive biomedical data.	5	E	1, 6
29	A generative AI company deploys a large language model capable of generating realistic news articles. Discuss two imminent and one long-term risk from this deployment, and propose one governance measure to manage each.	5	E	1, 6
30	A smart city surveillance system uses facial recognition for crime prevention but is criticized for privacy violations. Evaluate how privacy-by-design and data minimization could be integrated to balance safety and privacy.	5	E	1, 5

31	A healthcare chatbot trained on global data provides incorrect recommendations for local Indian populations. Identify sources of bias, their ethical implications, and recommend dataset or model adjustments for fairness.	5	AN	1, 2
32	A machine learning model used for college admissions predicts student success but performs poorly for first-generation learners. Examine the ethical and fairness issues involved and outline steps to audit and mitigate bias.	5	E	2, 4
33	A tech firm deploys a content moderation AI that flags minority languages more often as "offensive." Explain how interpretability tools (like LIME/SHAP) can help diagnose the issue and improve model fairness.	5	AP	3, 4
34	A large e-commerce company uses AI to personalize prices for users. Investigate potential ethical concerns and propose responsible AI guidelines to ensure consumer trust and fairness.	5	AN	1, 5
35	A multinational corporation deploys a multilingual customer support chatbot powered by a large language model. Users report culturally insensitive or biased responses in certain languages. Analyze the sources of such bias and recommend technical and organizational interventions to ensure inclusive and responsible AI behavior.	5	AN	1, 2, 5
36	Provide an example of how an AI system in the legal domain (e.g., predictive policing, bail risk assessment) could introduce bias. Explain the risks and suggest measures to ensure fairness and accountability.	5	AN	CO1, CO2
37	Using a real-world scenario, explain how AI in healthcare (e.g., diagnosis, treatment recommendation) can be unfair to minority groups. What interventions could improve fairness and reliability?	5	E	CO2, CO4
38	Give an example of how a student performance prediction system in education might misinterpret data, leading to biased outcomes. Explain methods to audit and correct such biases.	5	AP	CO2, CO5
39	Provide an example of a policy failure in AI deployment (e.g., facial recognition in public spaces). Explain how improved RAI principles could have prevented the issue.	5	E	CO1, CO6
40	Illustrate with an example how AI-assisted legal tools (e.g., contract review AI) might produce errors. Explain how audit mechanisms or human oversight can mitigate risks.	5	AN	CO3, CO5

Criteria	**Excellent(100%)**	**Good(75%)**	**Fair(50%)**	**Poor(25% or below)**
Identification of risks (1.5), Relevance of examples (1.5)	Identifies 3 risks and gives 3 well-matched examples	Identifies 2–3 risks, with 2 partial examples	Identifies 1–2 risks, vague or unclear examples	Incorrect or missing risks, no valid example
Accurate definition (1), Clear example (1)	Clear definition and relevant harmful behavior example	Correct definition, general example	Partial or vague definition, poor example	Incorrect or no definition, no example
Definitions (2), Relevant examples (2)	Correctly defines all 3 risks and gives strong examples	Defines 2–3 risks, examples moderately relevant	Defines 1–2 risks, examples weak or unclear	Incorrect definitions, examples missing or wrong
Explanation of principles (2), Ethical relevance (2)	Explains 4 valid principles, shows relevance to ethics/inclusion	Explains 3–4 principles, moderate ethical connection	Explains 2–3 principles, relevance weak	Less than 2 principles or vague/irrelevant content
Definition (1.5), Comparison (2.5)	Correct definition + strong comparison of vision vs NLP attacks	Definition clear + comparison partially explained	Vague definition + little or unclear comparison	Incorrect/no definition, no comparison
Interpretability definition (1.5), Model advancement examples (1.5)	Clear definition + 2 valid, recent advancements	Clear definition + 1–2 partially relevant advancements	Vague definition + outdated/unclear examples	Incorrect or no definition, no examples